

# The Reading the Mind in the Eyes Test: Test-retest Reliability and Preliminary Psychometric Properties of the German Version

Monique C. Pfaltz<sup>\*1</sup>, Salome McAleese<sup>2</sup>, Andreas Saladin<sup>3</sup>, Andrea Hans Meyer<sup>4</sup>, Markus Stoecklin<sup>5</sup>, Klaus Opwis<sup>5</sup>, Gerhard Dammann<sup>6</sup> and Chantal Martin-Soelch<sup>7,8</sup>

<sup>1</sup>Department of Psychology, Harvard University, Cambridge, USA

<sup>2</sup>Hoffmann-La Roche, Research and Early Development, Basel, Switzerland

<sup>3</sup>SAM Consulting Group, organizational development und change management, Basel, Switzerland

<sup>4</sup>Department of Clinical Psychology and Epidemiology, University of Basel, Basel, Switzerland

<sup>5</sup>Department of Cognitive Psychology and Methodology, University of Basel, Basel, Switzerland

<sup>6</sup>Psychiatric Clinic Muensterlingen, Muensterlingen, Switzerland

<sup>7</sup>Division of Clinical Psychology, Department of Psychology, University of Fribourg, Fribourg, Switzerland

<sup>8</sup>Department of Psychiatry and Psychotherapy, University Hospital Zurich, Zurich, Switzerland

<sup>\*</sup>pfaltz@wjh.harvard.edu

## Abstract

The Reading the Mind in the Eyes test (short Eyes test) is a widely used instrument assessing theory of mind abilities in adults. The present study for the first time assesses its test-retest reliability and provides initial data on the psychometric properties of a German version. 132 nonclinical participants completed the German Eyes test, a test of facial emotion recognition, and a measure of verbal skills. 40 of the 132 participants completed the Eyes test twice, three weeks apart. Results suggest that overall, the German Eyes test is a reliable instrument. No systematic learning effects occurred with repeated testing and measurement precision was evenly distributed across different ranges of performance. Moreover, a significant correlation between Eyes test scores and a related construct, the Facially Expressed Emotion Labeling (FEEL) test, supports the construct validity of the German translation. However, analyses of individual items (item difficulty, test-retest agreement) suggest that psychometric properties of certain items could be improved. Examining the psychometric qualities and clinical usefulness of a short version might thus prove fruitful. Furthermore, future research should assess whether the clinical strengths of the original version (in particular, the differentiation between individuals with autism spectrum disorders and nonclinical controls) also apply to the German version.

## Keywords

*Theory of Mind; Reading the Mind in the Eyes; Reliability;*

*German; Autism; Asperger*

## Introduction

Having a “theory of mind” is a crucial aspect of social cognition that refers to the ability to attribute mental states to oneself and others in order to understand and explain behaviour (A. Gopnik, 1997). Awareness of own and other people’s beliefs, perceptions, and emotions is not only essential for successful daily functioning in adults but can already be observed in 15-month toddlers (K. H. Onishi, 2005), underlining the importance of theory of mind abilities for social interactions even at a very young age. In the last decades, imaging studies demonstrated that theory of mind can be distinguished from other cognitive functions and is related to specific neural structures, including structures of the temporal lobe and the prefrontal cortex (T. Singer, 2006). Also, using theory of mind abilities is accompanied by amygdala activity in nonclinical controls but not in people with autism (S. Baron-Cohen, 1999), indicating a biological foundation for this cognitive ability.

Theory of mind deficits have been observed in different neurological and psychiatric conditions such as autism (S. Baron-Cohen, 2001), schizophrenia (M. Bruene, 2005), and bipolar disorder [(N. Kerr, 2003), (L.

S. Schenkel, 2008)]. To examine theory of mind abilities in clinical research and practice, the development of adequate instruments is essential. For children, various social cognitive tests have been developed (M. Sprung, 2010). However, sensitive instruments assessing theory of mind deficits in adults with otherwise normal intelligence are rare. They are difficult to develop since adults suffering from conditions like autism or Asperger syndrome may have acquired compensatory strategies that cover their deficits in social cognition to a certain degree, resulting in subtle deficits that are difficult to detect. Thus, sensitive tests are needed to detect subtle theory of mind deficits. The "Reading the Mind in the Eyes Test" has been developed to provide a sensitive test for the assessment of theory of mind abilities in adults [(S. Baron-Cohen, 1997), (S. Baron-Cohen, 2001)]. The Eyes test became an established instrument in the domain of social cognition. It involves theory of mind abilities in the sense that participants have to understand mental state terms and match them to the eye region of faces. Since the publication of an elaborated version that became the standard version of the test and detects subtle deficits in individuals with Asperger syndrome, high-functioning autism and controls (S. Baron-Cohen, 2001), the Eyes test has been translated to various languages (e.g. Japanese, French, and Spanish). It is widely used in clinical practice and scientific studies but research assessing its psychometric properties, in particular the properties of translated versions, is severely lacking. In German speaking countries, the adult Eyes Test has been used in several studies, [(S. Baron-Cohen, 2001), (G. Nietlisbach, 2010), (M. Voracek, 2006)], yet studies focusing on its psychometric properties are lacking, and the translation is not standardized.

The aim of this study was to provide preliminary data on the psychometric properties of a German version of the Adult Eyes test in a nonclinical sample, following a methodology similar to the one used for the development of the original test (S. Baron-Cohen, 2001). In addition, we aimed to assess the test-retest reliability of the German Eyes test. The test-retest reliability of the Eyes test has only been assessed for the Swedish child version, which shows a fairly high test-retest reliability (M. U. Hallerbäck, 2009), but not for the English original (S. Baron-Cohen, 2001) or previously used German translations [(S. Baron-Cohen, 2001), (G. Nietlisbach, 2010), (M. Voracek, 2006)]. We therefore examined if the findings for the Swedish child version (M. U. Hallerbäck, 2009) can be

replicated for the German adult Eyes test. Furthermore, we carried out the following, subsidiary analyses: Firstly, in addition to the Eyes test, we used the Facially Expressed Emotion Labeling (FEEL) test (H. Kessler, 2002), a computer based assessment of the ability to recognize basic facial emotions. The FEEL test served to examine the construct validity of the German Eyes Test by assessing its association with a related construct. Secondly, we assessed a potential influence of verbal ability on peoples' performance in the Eyes test since verbal skills predicted results of the Eyes Test in healthy adults (F. S. Ahmed, 2011), and in adults with Asperger syndrome and high-functioning autism, verbal IQ had a significant impact on Eyes test results (O. Golan, 2006). However, this influence was not confirmed in other studies (A. Senju, 2002). Finally, we assessed whether women show a better Eyes test performance than men, which would confirm previous findings [(S. Baron-Cohen, 1997), (S. Baron-Cohen, 2001), (M. U. Hallerbäck, 2009)].

## Methods

### *Participants*

155 participants were recruited among students at the University of Basel (Switzerland) and from the general population in the region of Basel by providing study information at adult education centers and public talks held by the senior author. Eighty-six (55.8%) of the participants were female. Mean age was 31.2 (SD=12.7). One hundred and ten (71%) of the participants had completed high school. For twenty (12.9%) participants, the highest level of completed education was a university degree (Masters or PhD) fourteen (9%) had completed commercial school, four had (2.6%) an apprenticeship, three had (1.9%) a higher education college degree, and four had (2.6%) another education.

### *Materials*

*Eyes test.* The original Eyes test has been developed in 1997 (S. Baron-Cohen). In this version, participants were presented 25 pictures of the eyes regions of the face, and asked to choose between two words that best describe the thoughts and feelings of the person in the picture. At a rapid and subconscious level, people are assumed to sort and match the eyes to faces stored in memory, and make according judgments regarding the mental states of the faces (S. Baron-Cohen, 2001). In 2001, an elaborated version with improved psychometric properties was published (S. Baron-Cohen, 2001). This version includes 36 items and for

each picture, participants have to choose one of four mental state terms. Here, we tested the psychometric properties of the revised version. The test was first translated from English to German by a bilingual undergraduate student in German literature. The German version was then back-translated to English by a native English-speaker with very good knowledge of German. German foils and target mental state terms are listed in Appendix A. As in the English version (S. Baron-Cohen, 2001), participants additionally had to decide whether the presented eyes belonged to a man or a woman. This task served as control measure of basic social face perception.

*Basic Emotion Recognition Abilities.* To assess the construct validity of the German Eyes test, we examined its association with the Facially Expressed Emotion Labeling (FEEL) test (H. Kessler, P. Bayerl, 2002), which assessed basic emotion recognition abilities. The FEEL test is a validated computerized test consisting of 42 photographs of people's faces appearing briefly on a screen. After a short delay, participants have to decide which of the six basic emotions (happiness, sadness, disgust, fear, surprise, anger) was presented. The validity of the FEEL test is supported by negative associations with scores of the German version of the Toronto Alexithymia Scale [TAS; (J. Kupfer, 2000)], and in particular with the TAS scale "difficulty recognizing feelings".

*Verbal skills.* The Mehrfachwahl-Wortschatz-Intelligenztest (MWT), a multiple choice word comprehension test with high reliability (Parallel test correlation between version A and B:  $r=0.84$ , test-retest reliability:  $r=0.87$  after 14 months) [MWT-A; (S. Lehl, 1974), (S. Lehl, 1995)], served as an index of verbal skills. The MWT is a widely used German test of semantic knowledge and language related intelligence (J. L. Horn, 1966). It can be considered a German equivalent to the Spot-The-Word test (A. Baddeley, 1993). Each of its 37 items comprises one word and four fictive words (e.g. "nedarid") and participants have to identify the real word. A verbal intelligence quotient (IQ) can be calculated, serving as measure of verbal skills in this study.

### **Procedure**

The study was approved by the local ethics committee for medical research. After participants gave written informed consent, the Eyes Test was administered in a quiet room of the University of Basel. Participants were instructed to look carefully at each picture and -

after reading the four words - to indicate as fast as possible which word describes best what the person whose eyes are shown is feeling or thinking. They could use a glossary at any point during testing to look up word meanings. Before completing the Eyes Test, 132 of the 155 participants (74 women) carried out the MWT (S. Lehl, 1974) and the FEEL Test (H. Kessler, 2002). For the remaining participants, these data could not be collected. To assess the test-retest reliability of the German Eyes Test, 40 of the 132 participants (21 women), completed the test twice (average between testings: 3 weeks). Participants received no compensation but a written report regarding their performance on the completed tests.

### **Statistical Analyses**

Statistical analyses were performed using SPSS 18.0 (SPSS Inc, Chicago, Illinois). Eyes test scores were built by calculating the sum across all items or across items considered acceptable based on item difficulty and test-retest analyses (see below). Due to right-skewed distributions of FEEL and Eyes test scores, Mann Whitney U-tests were used to evaluate gender differences in FEEL and Eyes test scores.

Analyses that only included Eyes test scores of the first assessment were based on the total sample ( $n=155$ ). Analyses that included data of the second Eyes test assessment were based on the subsample of 40 participants completing the Eyes test twice. Remaining analyses included the data of the 132 participants who, in addition to completing the Eyes test, had completed the MWT and FEEL test. For one participant, gender information was unavailable. Therefore, gender differences in Eyes test scores were assessed in  $n=154$  participants. For one participant, data of the MWT and for two participants, data of the FEEL-test could not be analyzed due to technical difficulties. To assess associations between Eyes and FEEL test scores and between Eyes test scores and verbal IQ, Spearman's rank correlation coefficient was used. For all statistical analyses, alpha was set to 0.05.

*Test-Retest Reliability.* Associations between Eyes test scores of the first and second assessment were calculated using Spearman's rank correlation coefficient. In addition, following the approach used for the assessment of the test-retest reliability of the Swedish child version (H. Kessler, 2009), we used the Bland-Altman method (J. M. Bland, 1986). This method measures agreement between two instruments or between the same instruments at different assessment points. The Bland-Altman plot in Figure 1

visualizes the agreement between Eyes test scores of the two assessments by plotting the difference between test and retest scores against the mean of test and retest scores for each participant. The mean and standard deviation of the differences assess the lack of agreement between assessments. Confidence intervals for the mean difference are calculated to determine if the mean difference deviates significantly from zero. Finally, upper and lower limits of agreement, indicating the range within which 95% of the test scores of two assessments can be expected to vary, are drawn.

Retest reliability of the Eyes test was also assessed for each item separately by calculating the percentage agreement (proportion of cases in which participants selected either target or foil at both time points). Following the procedure used in (T. L. Hutchins, 2008), an agreement value of at least 70% was considered as criterion for acceptable retest reliability.

## RESULTS

### *Gender Recognition Cotrol Task*

One participant identified the gender correctly only for 47% of the items. The Eyes test score of this participant was however in the normal range. Therefore, her data were included in the analyses. On average, the remaining participants were able to differentiate correctly between male and female eyes in 95% of the cases (range: 78%-100%).

### *Item Difficulty*

Table 1 shows the percentage of participants who chose each word on each item, which served as index of item difficulty. Based on used criteria during test development (S. Baron-Cohen, 2001), items were considered to be of satisfying item difficulty if at least 50% selected the target word and if no more than 25% selected one of the foils. On five items (2, 7, 21, 25, 31), the target word was selected by fewer than 50% of the participants. On six items (2, 13, 17, 21, 25, 31), one of the foils was selected by more than 25%.

### *Test-retest Reliability*

Table 2 shows the percentage of agreement between test administrations. Based on our criterion of an agreement value of at least 70%, items 7, 8, 13, 14, 16, 17, 21, 29, 31, and 32 did not achieve acceptable reliability. Together with the findings regarding item difficulty, these results suggest that items 2, 7, 8, 13, 14, 16, 17, 21, 25, 29, 31, and 32 display questionable psychometric properties. Therefore, the remaining

results will be presented separately for all 36 items and for the 24 items considered to be acceptable.

Eyes test scores of the first and second session were significantly positively correlated ( $r = 0.68$ ,  $p < 0.001$  for all items;  $r = 0.66$ ,  $p < 0.001$ , for the 24 acceptable items). Results of the Bland-Altman approach showed a mean difference of  $-0.78$  ( $SD = 3.09$ ) between test scores (see Figure 1). When excluding items with questionable psychometric properties, the mean difference was  $-0.43$  ( $SD = 2.58$ ). The 95% confidence interval for the mean difference was  $-1.77$  to  $0.20$  ( $-1.26$  to  $0.39$  when excluding items with questionable psychometric properties). Upper and lower limits of agreement were  $5.40$  and  $-6.96$  ( $4.73$  and  $-5.59$  when excluding items with questionable psychometric properties; see Figure 1). Thus, when considering the mean difference as 0, the limits of agreement, within which 95% of the test scores of two assessments can be expected to vary, can be described as  $\pm 6.18$  ( $\pm 5.16$  when excluding items with questionable psychometric properties).

TABLE 1. PERCENTAGES OF PARTICIPANTS WHO CHOSE EACH WORD ON EACH ITEM (N=155)

Item	Target	Foil 1	Foil 2	Foil 3
1	65.8	17.4	14.8	1.9
2 <sup>a</sup>	49.4	33.8	1.9	14.9
3	85.1	3.2	0.6	11.0
4	74.2	2.6	7.7	15.5
5	64.5	16.8	18.1	0.6
6	72.9	7.7	11.6	7.7
7 <sup>a</sup>	49.0	12.9	22.6	15.5
8 <sup>a</sup>	77.4	1.9	3.9	16.8
9	78.6	3.2	15.6	2.6
10	76.0	15.6	6.5	1.9
11	74.3	13.2	5.9	6.6
12	87.7	4.5	4.5	3.2
13 <sup>a</sup>	55.8	11.7	6.5	26.0
14 <sup>a</sup>	73.4	15.6	7.8	3.2
15	84.5	0.0	1.9	13.5
16 <sup>a</sup>	76.0	5.2	2.6	16.2
17 <sup>a</sup>	50.3	40.6	0.6	8.4
18	81.9	5.8	4.5	7.7
19	57.4	14.8	20.6	7.1
20	81.3	9.0	8.4	1.3
21 <sup>a</sup>	39.4	52.9	7.1	0.6
22	72.9	3.2	12.3	11.6
23	61.7	3.9	11.0	23.4
24	57.4	11.6	14.8	16.1
25 <sup>a</sup>	42.6	0.0	48.4	9.0
26	78.1	12.3	5.8	3.9
27	67.1	1.3	17.4	14.2
28	63.9	5.2	22.6	8.4
29 <sup>a</sup>	69.0	8.4	6.5	16.1
30	86.5	1.3	5.2	7.1
31 <sup>a</sup>	32.3	21.3	16.8	29.7
32 <sup>a</sup>	66.5	1.9	10.3	21.3
33	77.4	6.5	14.2	1.9
34	71.0	3.2	10.3	15.5
35	60.6	16.8	12.9	9.7
36	85.8	1.3	2.6	10.3

a. Items with questionable psychometric properties.

TABLE 2. PERCENTAGE OF AGREEMENT BETWEEN TEST ADMINISTRATIONS (N=40).

Item	Percentage agreement	Item	Percentage agreement
1	90	19	75
2	77	20	80
3	78	21 <sup>a</sup>	53
4	85	22	80
5	73	23	73
6	80	24	75
7 <sup>a</sup>	68	25	85
8 <sup>a</sup>	69	26	73
9	85	27	73
10	73	28	80
11	74	29 <sup>a</sup>	65
12	77	30	88
13 <sup>a</sup>	70	31 <sup>a</sup>	60
14 <sup>a</sup>	68	32 <sup>a</sup>	60
15	93	33	73
16 <sup>a</sup>	69	34	79
17 <sup>a</sup>	65	35	77
18	88	36	88

<sup>a</sup> Percentage agreement <70.

TABLE 3. MEAN (SD) AND MEDIAN EYES AND FEEL TEST SCORES AND

		Total Group	Males	Females
Eyes test (all items, n=36)	Mean (SD)	24.5 (3.5)	24.8 (3.2)	24.2 (3.7)
	Median	25.0	24.5	25.0
	Range	13–31	16–31	13–31
	n	155	68	86
Eyes test (problematic items excluded, n=24)	Mean (SD)	17.7 (2.8)	17.8 (2.7)	17.5 (2.9)
	Median	18.0	18.0	18.0
	Range	8–24	11–23	8–24
	n	155	68	86
FEEL test	Mean (SD)	36.1 (3.7)	36.7 (3.2)	35.7 (4.0)
	Median	37.0	37.0	36.0
	Range	25–42	25–42	25–42
	n	129	58	71
Verbal IQ	Mean (SD)	123.5 (12.1)	124.5 (11.3)	122.7 (12.8)
	Range	80–143	98–143	80–143
	n	131	58	73

VERBAL IQ FOR MALES, FEMALES, AND THE TOTAL STUDY GROUP.

### Mean Eyes Test Scores

Mean (SD) and median scores of correct answers out of 36 and 24 items are displayed in Table 3, separately for males, females, and the total study group.

### Correlations with FEEL Test and Verbal IQ

Table 3 gives descriptive statistics for FEEL test and verbal IQ. Correlations between FEEL and Eyes test scores were significant, regardless of whether items with questionable psychometric properties were included ( $r = 0.43$ ,  $p < .001$ ) or excluded ( $r = 0.43$ ,  $p < .001$ ). When excluding ( $r = 0.173$ ,  $p = .048$ ) but not when including ( $r = 0.095$ ,  $p = .282$ ) items with questionable psychometric properties, the correlation between verbal IQ and Eyes test scores reached significance.

### Gender Differences

Males did not show lower Eyes test scores than females ( $U = 2671$ ,  $p = 0.355$ ;  $U = 2754$ ,  $p = 0.533$  when excluding items with questionable psychometric properties). No gender difference in FEEL test scores occurred ( $U = 1813$ ,  $p = 0.242$ ).

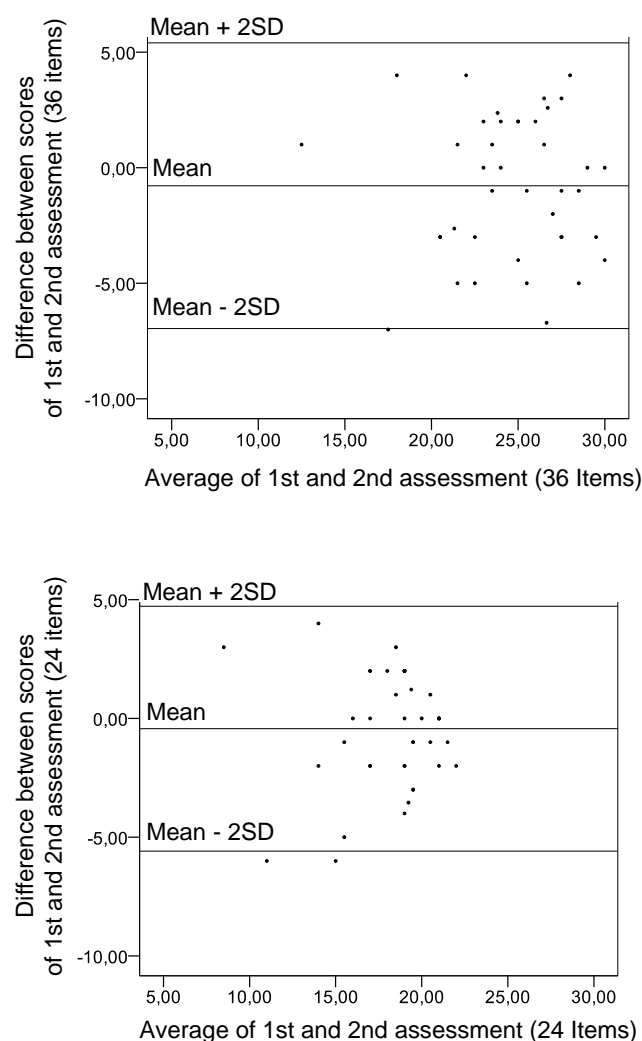


FIGURE 1. AGREEMENT BETWEEN THE TWO EYES TEST ASSESSMENTS (N=40)

## Discussion

The present study assessed the test-retest reliability of the German Eyes test and provides initial data on its psychometric properties. Our sample showed somewhat lower test results compared to the English version (S. Baron-Cohen, 2001). The German translation might have entailed subtle alterations in the meaning of targets and foils, which may have increased the difficulty of certain items, resulting in lower test scores. In fact, for several items, correct answers were not selected by a clear majority of participants (Table 1) and thus need to be considered too difficult. Moreover, with respect to percentage agreement values, some of these items and several additional items did not achieve acceptable reliability, resulting in 12 out of 36 items with questionable psychometric properties.

While percentage agreement values and findings regarding item difficulty suggest that on the individual item level, psychometric properties of the German Eyes test could be improved, results of the Bland-Altman reliability analysis, a method assessing the range within which test scores of two assessments can be expected to vary, are generally satisfactory. Differences in test scores between assessments did not vary systematically with lower or higher test scores (see Table 1), pointing to an even distribution of measurement precision across different ranges of performance. Furthermore, zero was included in the confidence interval for the mean difference between first and second assessment, and differences in test scores were equally distributed in both directions around zero. Thus, no systematic learning effects seem to occur with repeated usage of the German Eyes test. Despite these promising results, it should be taken into account that individual results can vary in the range of  $\pm 6$  out of 36 or  $\pm 5$  out of 24 possible with repeated assessments. This is comparable to the Swedish child version, for which a variation of  $\pm 4$  out of 24 possible was found (M. U. Hallerbäck, 2009). Nevertheless, the variation we found for repeated test administrations shows that although the Eyes Test has been widely spread, further evaluations of its psychometric properties are needed. In fact, this is the first study assessing test-retest reliability for the adult version of the Eyes test. It thus remains to be examined if the original version and other widely used translations include items with poor retest-reliability. Also, future studies should assess intra-individual

differences with repeated testing in relation to differences in Eyes test scores between different clinical and non-clinical groups.

## *Associations with FEEL Test and Verbal Skills*

We found a moderate but significant correlation between Eyes and FEEL test scores, speaking for the construct validity of the German Eyes test. Eyes and FEEL test both require basic emotion recognition abilities. Thus, their results should be correlated to a certain degree. However, the FEEL test differs from the Eyes test since it provides information from the whole face and only covers the six basic emotions. Thus, the correlation between the Eyes test and a related, but not identical construct like the FEEL test should only be in the moderate range. Clearly, more research is needed on the construct validity of the German Eyes test. The original Eyes test correlates with two related constructs: The Empathy Quotient [(E. J. Lawrence, 2004), (S. Baron-Cohen, 2004)] and the Autism Spectrum Quotient (S. Baron-Cohen, 2001). Using German translations of the three tests, these findings were confirmed (M. Voracek, 2006). However, the correlations that were found were small and lower than the ones found for the original Eyes test [(S. Baron-Cohen, 2004), (S. Baron-Cohen, 2001)], pointing to potential problems with the convergent validity of the German Eyes test. These findings should however be interpreted with caution since they are based on not validated translations of these tests and use a slightly different methodology as in the validation study of the original English test (i.e. no presentation of the glossary of mental-state terms).

When excluding items with questionable psychometric properties, we found a small but significant correlation between Eyes test scores and verbal skills. Previous studies assessing this association found mixed results, which might be explained by the assessment of different populations by varying measures of verbal skills. For instance, Senju and colleagues (A. Senju, Y. Tojo, 2002) estimated verbal IQ by subscales of the Wechsler Intelligence Scale for Children-Revised (D. Wechsler, 1974) and found no association with Eyes test scores in children with autism. In healthy adults, Eyes test scores were unrelated to tests of verbal fluency and verbal abstraction but were predicted by the Wechsler Test of Adult Reading [(The Psychological Corporation, 2010), (M. Voracek, 2006)]. Using the Wechsler Abbreviated Scale of Intelligence (D. Wechsler, 1999), others (O.

Golan, 2006) found a significant impact of verbal IQ on Eyes test scores in adults with autism spectrum conditions and controls. Together with previous research [(F. S. Ahmed, 2011), (O. Golan, 2006)] our findings suggest that next to the ability to decode social information, verbal skills might contribute to successful Eyes test performance. However, having the opportunity to look up word meanings should reduce the likelihood that people who do not understand certain words answer the corresponding items incorrectly, possibly explaining our low effect size ( $r=0.17$ ) for the correlation between verbal skills and Eyes test scores.

### *Gender Differences*

Previously found gender differences in Eyes test scores [(S. Baron-Cohen, 1997), (S. Baron-Cohen, 2001), (M. U. Hallerbäck, 2009)] could not be replicated. Men did not perform significantly worse than women, neither on the Eyes nor on the FEEL test. We thus found no evidence for weaker basic emotion recognition or more complex theory of mind abilities in men. In two of the previous studies, Eyes score gender differences were small or only reached marginal significance [(S. Baron-Cohen, 2001), (M. U. Hallerbäck, 2009)]. The third study (S. Baron-Cohen, 1997) found a highly significant gender difference of three points but used the unrevised version of the Eyes test. Additional studies are needed to assess if the distinct female advantage found in (M. U. Hallerbäck, 2009) can be replicated for the revised version of the Eyes test and to clarify if cultural differences might explain discrepant findings.

### *Limitations*

This study has certain limitations. First, in the original Eyes test (S. Baron-Cohen, 2001), judges determined the correct answers. It is unclear if, due to cultural differences, German speaking judges would have arrived at different solutions. By adopting correct answers from the English original, we might erroneously have rated certain answers as incorrect. This could explain why, according to the English version, five out of 36 items were answered incorrectly by more than 50% of our sample, and why our sample reached lower scores compared to the English version (S. Baron-Cohen, 2001). If this was due to cultural differences, lower scores in our sample should not be attributed to worse performance. However, the use of the original set of answers improves the comparability

of our results with other translations of the test. Second, future studies should include more participants with lower education levels to provide reference values that are more representative for the general population. Furthermore, we did not assess participants' full scale intelligence, which might have influenced Eyes test performance. Previous findings regarding this potential influence are mixed. Some researchers [(S. Baron-Cohen, 2001), (O. Kelemen, 2004)] found no association between full scale intelligence and Eyes test performance, while others (M. Losh, 2006) found such an association in parents of children with autism. When excluding items with questionable psychometric properties, we found evidence for a small but significant association between verbal intelligence and Eyes test scores. It might thus prove fruitful to further assess the common and separate impact of verbal and full scale intelligence on peoples' theory of mind abilities. Finally, a further study should test the psychometric qualities of the test in a much larger sample.

### *Conclusion*

Our analysis of individual items suggests that psychometric properties of certain items could be improved. Results of the remaining reliability analysis are satisfactory and the correlation between Eyes and FEEL test is promising with respect to the construct validity of the German Eyes test. However, future studies should assess the validity of the German Eyes test in more detail. A next step will be to use a contrasting-groups method of construct validation to test if the German Eyes test is able to differentiate between nonclinical controls and individuals with autism-spectrum disorders. The distribution of Eyes test scores (Table 3) does not point to a ceiling effect and illustrates that scores of the German version cover a large range. This suggests that similarly to the English version (S. Baron-Cohen, 2001), distribution of test scores will leave room to differentiate between nonclinical and clinical study groups and to measure different degrees of the ability to interpret mental states from the eyes region (as an important aspect of theory of mind abilities) in disorders like autism or Asperger syndrome – hypotheses which should be tested. Next to assessing the validity of the German Eyes test in more detail, it might prove fruitful to validate and assess the clinical usefulness of a short version of the German Eyes test which excludes items for which we found questionable psychometric properties.

To conclude, our data provide initial evidence for some limitations of the psychometric qualities of the German Eyes test and call for more studies and possibly adapted versions. Despite certain psychometric problems we found, the Eyes test will remain an important clinical instrument, which, according to the original version (S. Baron-Cohen, 2001), is able to differentiate between nonclinical populations and individuals with autism spectrum disorders, underlining the clinical strengths of the test. Together with further assessments of its psychometric properties, the Eyes test and the present findings provide a basis for future investigations of disorders characterized by theory of mind deficits, and will enhance comparability of findings across countries.

#### ACKNOWLEDGMENT

We thank Kirsten Taylor for her help with the translation, Beatrice Moerstedt and Panagiota Mistridis for their help with data analyses, and Professor Baron-Cohen for providing the test material. We thank the Faculty of Psychology of the University Basel for funding this study.

#### REFERENCES

- Baddeley, H. Emslie and I. Nimmo-Smith, *Br. J. Clin. Psychol.* 32, 55–65 (1993).
- D. Wechsler, *Manual for the Wechsler Abbreviated Scale of intelligence*, Psychological Corporation, San Antonio (1999).
- D. Wechsler, *Wechsler Intelligence Scale for Children-Revised*, Psychological Corporation, New York (1974).
- E. J. Lawrence, P. Shaw, D. Baker, S. Baron-Cohen and A. S. David, *Psychol. Med.* 34, 911–924 (2004).
- F. S. Ahmed and S. Miller, *J. Autism Dev. Disord.* 41(5), 667–678 (2011).
- G. Domes, M. Heinrichs, A. Michel, C. Berger and S. Herpertz, *Biol. Psychiatry* 61, 731–733 (2007).
- G. Nietlisbach, A. Maercker, W. Rössler and H. Haker, *Psychol. Rep.* 106, 832–844 (2010).
- Golan and S. Baron-Cohen, *Dev. Psychopathol.* 18, 591–617 (2006).
- Gopnik and A. N. Meltzoff, *Words, Thoughts, and Theories*, MIT Press, Cambridge (1997).
- H. Kessler, P. Bayerl, R. M. Deighton and H. C. Traue, *Verhaltenstherapie und Verhaltensmedizin* 23(3), 297–306 (2002).
- J. Kupfer, B. Brosig and E. Braehler, *Z. Psychosom. Med. Psychother.* 46, 368–384 (2000).
- J. L. Horn and R. C. Cattell, *J. Educ. Psychol.* 57, 253–280 (1966).
- J. M. Bland and D. G. Altman, *Lancet* 1, 307–310 (1986).
- K. H. Onishi and R. Baillargeon, *Science* 308, 255–258 (2005).
- Kelemen, S. Keri, A. Must, G. Benedek and Z. Janka, *Acta psychiatry. Scand.* 110, 146–149 (2004).
- L. S. Schenkel, M. Marlow-O'Connor, M. Moss, J. A. Sweeney and M. N. Pavuluri, *Psychol. Med.* 38(6), 791–800 (2008).
- M. Bruene, *Schizophr. Bull.* 31, 21–42 (2005).
- M. Losh and J. Piven, *J. Child. Psychol. Psyc.* 48, 105–112 (2006).
- M. Sprung, *Child. Adolesc. Ment. Health.* 15, 204–206 (2010).
- M. U. Hallerbeck, T. Lugnegard, F. Hjorthag and C. Gillberg, *Cogn. Neuropsychiatry* 14(2), 127–143 (2009).
- M. Voracek and S. G. Dressler, *Pers. Individ. Dif.* 41, 1418–1491 (2006).
- N. Kerr, R. I. Dunbar and R. P. Bentall, *J. Affect. Disord.* 73, 253–259 (2003).
- S. Baron-Cohen and S. Wheelwright, *J. Autism. Dev. Disord.* 34, 163–175 (2004).
- S. Baron-Cohen, H. Ring, S. Wheelwright, E. Bullmore, M. Brammer, A. Simmons et al., *Eur. J. Neurosci.* 11, 1891–1898 (1999).
- S. Baron-Cohen, *Int. Rev. Res. Ment. Retard.* 23, 169–203 (2001).
- S. Baron-Cohen, S. Wheelwright, J. Hill, Y. Raste, and I. Plumb, *J. Child. Psychol. Psychiatr.* 42(2), 241–251 (2001).
- S. Baron-Cohen, S. Wheelwright, R. Skinner, J. Martin and I. Clubley, *J. Autism. Dev. Disord.* 31, 5–17 (2001).
- S. Baron-Cohen, T. Jolliffe, C. Mortimore and M. Robertson, *J. Child. Psychol. Psychiatr.* 38, 813–822 (1997).
- S. Lehrl, J. Merz, H. Erzigkeit and V. Galster, *Mehrfachwahl-Wortschatz-Test*, Spitta Verlag, Balingen (1974).
- S. Lehrl, *Mehrfachwahl-Wortschatz-Intelligenztest (MWT-B)*, Straube, Erlangen (1995).
- Senju, Y. Tojo, Y., M. Konno, H. Dairoku and T. Hasegawa, *Jpn. J. Psychol.* 73(1), 64–70 (2002).
- T. L. Hutchins, P. A. Prelock and W. Chace, *Focus Autism. Other Dev. Disabl.* 23(4), 195–206 (2008).
- T. Singer, *Neurosci. Biobehav. Rev.* 30, 855–863 (2006).
- The Psychological Corporation, *Wechsler test of adult reading*, Harcourt Brace & Company, San Antonio (2001).

**APPENDIX A**

List of targets (in italic) and foils for each item.

1	<i>lustig</i>	beruhigend	genervt	Gelangweilt
2 <sup>a</sup>	verängstigt	<i>bestürzt</i>	arrogant	verärgert
3	scherzend	ausser Fassung	<i>begehrnd</i>	Überzeugt
4	scherzend	darauf <i>bestehend</i>	amüsiert	Entspannt
5	genervt	sarkastisch	<i>besorgt</i>	Freundlich
6	entgeistert	<i>tagträumend</i>	ungeduldig	Alarmiert
7 <sup>a</sup>	entschuldigend	freundlich	<i>unruhig</i>	Entmutigt
8 <sup>a</sup>	<i>verzweifelt</i>	erleichtert	schüchtern	Aufgeregt
9	verärgert	feindselig	entsetzt	<i>geistesabwesend</i>
10	<i>vorsichtig</i>	darauf bestehend	gelangweilt	Entgeistert
11	verängstigt	amüsiert	<i>bedauernd</i>	Kokett
12	gleichgültig	verlegen	<i>skeptisch</i>	Entmutigt
13 <sup>a</sup>	entschieden	<i>vorausahnend</i>	drohend	Schüchtern
14 <sup>a</sup>	genervt	enttäuscht	deprimiert	<i>beschuldigend</i>
15	<i>besinnlich</i>	ausser Fassung	ermunternd	Amüsiert
16 <sup>a</sup>	genervt	<i>nachdenklich</i>	ermunternd	Mitfühlend
17 <sup>a</sup>	<i>bezweifelnd</i>	zärtlich	lustig	Entgeistert
18	<i>entschieden</i>	amüsiert	entgeistert	Gelangweilt
19	arrogant	dankbar	sarkastisch	<i>Zögerlich</i>
20	dominant	<i>freundlich</i>	schuldig	Entsetzt
21 <sup>a</sup>	verlegen	<i>tagträumend</i>	verwirrt	in Panik
22	<i>geistesabwesend</i>	dankbar	darauf bestehend	Flehentlich
23	zufrieden	entschuldigend	<i>aufsässig</i>	Neugierig
24	<i>nachsinnend</i>	genervt	aufgeregt	Feindselig
25 <sup>a</sup>	in Panik	ungläubig	verzweifelt	<i>interessiert</i>
26	alarmiert	schüchtern	<i>feindselig</i>	Ängstlich
27	scherzend	<i>vorsichtig</i>	arrogant	Versichernd
28	<i>interessiert</i>	scherzend	zärtlich	Zufrieden
29 <sup>a</sup>	ungeduldig	entgeistert	genervt	<i>Tiefsinnig</i>
30	dankbar	<i>kokett</i>	feindselig	Enttäuscht
31 <sup>a</sup>	beschämt	<i>zuversichtlich</i>	scherzend	Entmutigt
32 <sup>a</sup>	<i>ernst</i>	beschämt	durcheinander	Alarmiert
33	verlegen	schuldig	<i>tagträumend</i>	<i>Beunruhigt</i>
34	entgeistert	verblüfft	<i>misstrauisch</i>	Verängstigt
35	verdutzt	<i>nervös</i>	darauf bestehend	Besinnlich
36	beschämt	nervös	<i>argwöhnisch</i>	unentschlossen